IRIA
ENTRE DE ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P 105
78153 Le Chesnay Cedex
France
Tél 954 90 20

Rapports de Recherche

N° 157

INVERSIONS EN CLASSIFICATION HIÉRARCHIQUE
APPLICATION À LA CONSTRUCTION ADAPTATIVE D'INDICES D'AGRÉGATION

Edwin DIDAY

Septembre 1982

INVERSIONS EN CLASSIFICATION HIERARCHIQUE :

Application à la construction adaptative d'indices d'agrégation

EDWIN DIDAY

Résumé

Une inversion se produit dans une hiérarchie quand l'indice associé à un palier h est supérieur à l'indice d'un palier h' bien que h soit inclus dans h'. Nous donnons :

1) Des conditions nécessaires et suffisantes sur les coefficients de la formule de récurrence permettant d'assurer l'existence ou l'inexistence d'inversions ; l'un de ces nouveaux résultats généralise les conditions données par Ducimetière (1971), Milligan (1979) et Batagelj (1981).
2) Une condition nécessaire et suffisante sur les même paramètres qui assure la propriété dite de "réductibilité" de Bruynooghe (1978) et permet ainsi d'utiliser des algorithmes accélérés de classification hiérarchique.

Nous montrons enfin comment on peut utiliser ces résultats pour étendre la famille des indices d'agrégation classiques de façon à mieux tenir compte du désir de l'utilisateur.

INVERSION HIERARCHICAL CLUSTERING :

adaptive ultrametrics

Abstract

We say that there is an inversion in a hierarchy if the height of a cluster merged later in the hierarchy is lower than that of clusters merged earlier. We give :

1) Necessary and sufficient conditions on the parameters of the Lance and Williams formula (1966), generalized by Jambu (1978), which insure the existence or non existence of inversion in a hierarchy ; one of those new results generalizes the conditions given by Ducimetière (1971), Milligan (1979) and Batagelj (1981).
2) A necessary and sufficient condition on the same parameters which insures the "neighbourhood reductibility" property of Bruynooghe (1978) and thus permits the use of an accelerated algorithm for hierarchical clustering.

By using those results and taking account of the a priori knowledge of the user we show that it is possible to obtain adaptive ultrametrics.

INVERSIONS EN CLASSIFICATION HIERARCHIQUE :

Application à la construction adaptative d'indices d'agrégation.

EDWIN DIDAY

1) Introduction

Il se produit fréquemment dans la pratique qu'un utilisateur désirant réaliser une classification hiérarchique ne sache pas quel sera le meilleur indice d'agrégation pour ses données parmi la panoplie usuelle des indices d'agrégation classiques ; il peut aussi se produire qu'aucun de ces indices ne satisfasse aux données qu'il doit traiter ; il se pose donc un problème de choix parmi les indices connus et un problème de création éventuelle de nouveaux indices. Parfois les données fournissent un indice naturel, par exemple le flux de transfert démographique d'un département à un autre si les individus à classer sont les départements. Malheureusement un tel indice peut faire apparaître des inversions au cours de la construction de la hiérarchie, ce qui la rend difficilement interprétable ; une inversion se produit quand l'indice associé à un palier h est supérieur à l'indice d'un palier h' bien que h soit inclus dans h'. Nous donnons des règles simples permettant de déduire de la valeur des coefficients de la formule de récurrence de Lance et Williams (1967) généralisée par Jambu (1978) l'existence possible d'inversions. Ces règles permettent également de savoir si la propriété de réductibilité définie par Bruynooghe (1978) est satisfaite et s'il est donc possible d'utiliser un algorithme de classification hiérarchique accélérée sans risque de déformation.
Si l'utilisateur a des idées a priori sur la hiérarchie qu'il désire obtenir, la formule de récurrence et les contraintes obtenues pour assurer l'inexistence d'inversion donne des équations dont les inconnues sont les coefficients $a_1 \ldots a_7$ de la formule de récurrence. S'il existe une solution elle fournit un indice d'agrégation adapté au désir de l'utilisateur.

INRIA, Domaine de Voluceau, Rocquencourt, 78153 LE CHESNAY CEDEX (France).

2) Quelques définitions

Définition d'une hiérarchie

Soit Ω un ensemble fini, H un ensemble de parties (appelées paliers) de Ω, H est une hiérarchie sur Ω si :

1) $\Omega \in H$ {c'est-à-dire le palier le plus haut contient tous les individus}
2) $\forall w \in \Omega$ $\{w\} \in H$ (les points terminaux)
3) $\forall h, h' \in H$ on a : $h \cap h' \neq \emptyset \Rightarrow h \subset h'$ ou $h' \subset h$.

Hiérarchie binaire

On appelle ainsi une hiérarchie H dont chaque palier est formé du regroupement de deux éléments de H , plus précisément :

$\forall B \in H : \text{card}(B) \geq 1, \exists (B', B'') \in H \times H :$
$B' \cap B'' = \emptyset$ et $B' \cup B'' = B$.

Définition d'une hiérarchie indicée

Une hiérarchie indicée est un couple (H,f) où H est une hiérarchie et f une application de H dans $\mathbb{R}^+$ telle que :

1) $f(h) = 0$ si et seulement si h ne contient qu'un seul élément.
2) pour tout h et h' dans H, $h \subset h'$ (inclusion stricte) implique $f(h) < f(h')$.

Autres types de hiérarchies indicées

Dans la pratique, certains algorithmes de classification hiérarchique (par exemple , celui de la CAH, qui est donné en annexe 1, (voir [2] ou [4]), peuvent donner lieu à des hiérarchies dont l'indice associé ne satisfait pas exactement la définition que nous venons de donner : il peut en effet se produire que deux éléments h et h' de H satisfassent aux relations $h \subset h'$, $h \neq h'$ et $f(h) = f(h')$. Dans ce cas, nous dirons que la hiérarchie est indicée au "sens large".

Définition d'une hiérarchie indicée au sens large

C'est un couple (H,f) où H est une hiérarchie et f une application de H dans $\mathbb{R}^+$ telle que :

1) f(h) = 0 si et seulement si h ne contient qu'un seul éléments.
2) pour tout h et h' dans H, $h \subset h'$ implique $f(h) \leq f(h')$.

Définition d'une inversion

Il peut aussi se produire que l'indice associé à certaines hiérarchies donne lieu à l'existence de paliers h et h' tels que $h \subset h'$ et $f(h) > f(h')$, on dit alors qu'il y a inversion ; pour une hiérarchie indicée ou pour une hiérarchie indicée au sens large, il n'y a pas d'inversions.

La construction d'une hiérarchie nécessite la connaissance d'une "mesure de ressemblance" entre groupes, cette "mesure" est appelée "indice d'agrégation".

Définition d'un indice d'agrégation

On appelle indice d'agrégation entre groupes d'individus, une application δ : $\mathbb{P}(\Omega) \times \mathbb{P}(\Omega) \to \mathbb{R}$ (où $\mathbb{P}(\Omega)$ est l'ensemble des parties de Ω) telle que :

1) $\forall h_1, h_2 \in \mathbb{P}(\Omega)$, $\delta(h_1,h_2) \geq 0$ (positivité)
2) $\forall (h_1,h_2) \in \mathbb{P}(\Omega) \times \mathbb{P}(\Omega)$ $\delta(h_1,h_2) = \delta(h_2,h_1)$ (symétrie).

Considérons une hiérarchie binaire H et un indice d'agrégation entre classes ; pour indicer H et pour être assuré qu'il n'y aura pas d'inversion on peut définir f de la façon suivante :

$\forall h_i, h_j$ dans H, $f(h_i \cup h_j) = \text{Max}\ (\delta(h_i,h_j), f(h_i), f(h_j))$.

On obtient ainsi une hiérarchie indicée au sens large. Par contre, si l'on choisit f ainsi :

$\forall h_i, h_j$ dans H $f(h_i \cup h_j) = \delta(h_i,h_j)$,

on n'est pas assuré que (H,f) soit une hiérarchie indicée, même au sens large, il peut se produire des inversions.

3) Conditions à satisfaire par un indice d'agrégation δ pour qu'une hiérarchie indicée par f : $f(h_1 \cup h_2) = \delta(h_1, h_2)$ ne présente pas d'inversions

La représentation visuelle d'une hiérarchie H est plus facilement interprétable si elle est indicée de façon à ce que la hauteur de chaque palier corresponde à la valeur prise par l'indice d'agrégation pour les deux paliers qui l'ont formé. Autrement dit, si f est choisi à partir de δ de la façon suivante :

$$\left.\begin{array}{l} f : H \to \mathbb{R}^+ \text{ telle que } f(h) = \delta(h_1, h_2) \\ \text{pour tout } h_1,\ h_2 \text{ et avec } h_1 \cap h_2 = \emptyset \text{ et } h = h_1 \cup h_2 \text{ dans } H \end{array}\right\} \quad (1)$$

Un tel choix de f peut conduire à des inversions d'où l'intérêt de la proposition suivante qui permet d'énoncer deux conditions nécessaires et suffisantes à satisfaire par δ pour que ce ne soit pas le cas.

Soit H la hiérarchie obtenue à l'aide de l'algorithme de la C.A.H. (voir annexe 1) muni de l'indice d'agrégation δ et f l'application définie par (1). On note P_{h_i} la partition qui précède la formation de $h_i = h_{i-1} \cup h_{i-2}$ dans la déroulement de l'algorithme (voir figure 1) :

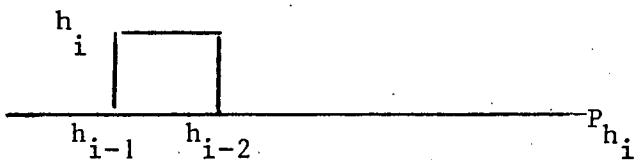


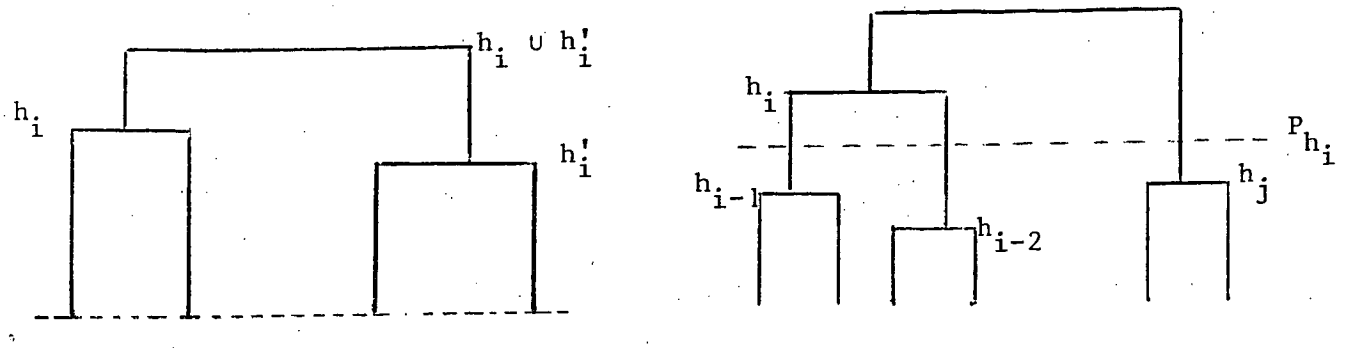
Figure 1

On peut alors énoncer le résultat suivant dont la démonstration se trouve dans [2] :

PROPOSITION 1. Les trois conditions suivantes sont équivalentes :

① (H,f) est une hiérarchie indicée au sens large.
② Pour tout $h_i \in H$, $h_i' \in H$ avec $h_i \cup h_i'$ dans H on a $f(h_i \cup h_i') \geq \text{Max}\{f(h_i), f(h_i')\}$.
③ Pour tout $h_j \in P_{h_i}$ avec $h_i = h_{i-1} \cup h_{i-2}$ et $h_j \neq h_{i-1}$, $h_j \neq h_{i-2}$ on a $\delta(h_i, h_j) \geq f(h_i)$.

Remarque pour fixer les idées, la condition ② peut être appelée "condition locale" et la condition ③ "condition globale" (voir figure 2) :



② Condition locale

$f(h_i \cup h'_i) \geq \text{Max}\ (f(h_i),\ f(h'_i))$

③ Condition globale

$\delta(h_i, h_j) \geq f(h_i)$

Figure 2

4) La formule de récurrence générale

Quand au cours de l'algorithme de la C.A.H. deux nouvelles classes h_1 et h_2 doivent être réunies pour former une classe $h_3 = h_2 \cup h_1 \in H$, il faut faire la mise à jour des $\delta(h_i, h_j)$ pour h_i et h_j dans P_{h_3} et différents de h_2 et h_1. Cette mise à jour nécessite le calcul de $\delta(h, h_1 \cup h_2)$ pour tout h dans P_{h_3} ; il se trouve que pour des choix classiques de l'indice d'agrégation δ, il est possible d'exprimer à l'aide d'une formule de récurrence $\delta(h, h_1 \cup h_2)$ à l'aide de $\delta(h, h_1)$, $\delta(h, h_2)$, $\delta(h_1, h_2)$, $f(h)$, $f(h_1)$ et $f(h_2)$ qui ont été calculés aux itérations précédentes ; ceci permet une grande économie de temps machine et de place mémoire ; on utilise pour cela la formule de récurrence générale suivante (Jambu 1978) :

$$\delta(h, h_1 \cup h_2) = a_1\, \delta(h,h_1) + a_2\, \delta(h,h_2) + a_3\, \delta(h_1,h_2) + a_4\ f(h) +$$

$$+ a_5\, f(h_1) + a_6\, f(h_2) + a_7 |\delta(h,h_2) - \delta(h,h_1)|$$

Cette formule généralise la formule classique de Lance et Williams (1967) qui ne considère pas le termes en a_4, a_5, a_6. Différents choix classiques des coefficients a_i sont donnés en annexe 2.

5) Une condition nécessaire et suffisante sur les coefficients de la formule de récurrence pour assurer l'inexistence d'inversions

Si la hiérarchie H est indicée de façon à ce que la hauteur de chaque palier corresponde à la valeur de l'indice d'agrégation pour les deux paliers qui l'ont formés. Autrement dit, si la condition suivante est satisfaite :

{pour tout h_1, h_2, h_3 dans H avec $h_3 = h_1 \cup h_2$ on a $f(h_3) = \delta(h_1, h_2)$} (2)

on peut se poser la question suivante :

Existe-t-il une condition nécessaire et suffisante sur les coefficients $(a_1, \ldots, a_7)$ pour assurer l'inexistence d'inversions dans la hiérarchie H indicée par f ?

La proposition suivante qui généralise la remarque de Ducimetière (1972), la proposition de Milligan (1979) et celle de Batagelj (1981) répond à la question.

PROPOSITION 2

Une condition nécessaire et suffisante pour que tout δ satisfaisant à une formule de récurrence définie par des coefficients $a_1, \ldots, a_7$, induise par l'algorithme de la CAH une hiérarchie indicée par f sans inversions est que les quatres conditions suivantes soient satisfaites :

(a) $a_7 \geq -\text{Min}\ (a_1, a_2)$

(b) $a_1 + a_2 \geq 0$

(c) $a_1 + a_2 + a_3 \geq 1$

(d) a_4, a_5 et a_6 positifs.

Démonstration

Pour simplifier les notations nous dirons que la condition A est vraie si les 4 relations (a), (b), (c), (d) sont simultanément satisfaites. La condition B est vraie si pour tout δ satisfaisant à la formule de récurrence définie par $a_1, \ldots, a_7$, l'algorithme de la CAH donne une hériarchie H indicée par f qui n'admet pas

d'inversions. Etant donné la symétrie de la formule de récurrence on peut supposer que $\delta(h,h_1) \leq \delta(h,h_2)$ (la démonstration est tout à fait analogue si cette inégalité a lieu dans le sens contraire).

Démontrons d'abord la condition suffisante, autrement dit que A vraie $\Longrightarrow$B vraie.

$\delta(h,h_1) \leq \delta(h,h_2)$ implique que :

$$\delta(h,h_1 \cup h_2) = (a_1 - a_7)\ \delta(h,h_1) + (a_2 + a_7)\ \delta(h,h_2) + a_3\ \delta(h_1,h_2) + a_4\ f(h) + a_5\ f(h_1) + a_6\ f(h_2).$$

(a) $\Rightarrow a_2 + a_7 \leq 0$

(d) $\Rightarrow a_4 \geq 0$, $a_5 \geq 0$, $a_6 \geq 0$ d'où

$$\delta(h,h_1 \cup h_2) \geq (a_1 + a_2)\ \delta(h,h_1) + a_3\ \delta(h_1,h_2)$$

Par construction de l'algorithme on a : $\delta(h,h_1) \geq \delta(h_1,h_2)$
et (b) $\Longrightarrow a_1 + a_2 \geq 0$ d'où

$$\delta(h,h_1 \cup h_2) \geq (a_1 + a_2 + a_3)\ \delta(h_1,h_2)$$

or (c) $\Rightarrow \delta(h,h_1 \cup h_2) \geq \delta(h_1,h_2)$ qui prouve que B est vraie d'après la proposition 1.

Montrons maintenant que B $\Longrightarrow$A ; plus précisement nous allons montrer que A faux $\Longrightarrow$B faux ; autrement dit, il s'agit de montrer que si A est faux, il existe un indice d'agrégation δ , pour lequel l'algorithme de la CAH donne une hiérarchie H indicée par f définie par (2), avec inversions.

Les quinze possibilités ($\sum_{n=1}^{4} C_4^n = 15$) qui rendent A faux (en terme logique non A = non $a \wedge b \wedge c \wedge d = \overline{A} = \bar{a} \vee \bar{b} \vee \bar{c} \vee \bar{d}$) sont couvertes par les quatre cas suivants :

① $a_2+a_7 < 0$ ou $a_1 + a_7 < 0$.

② $a_1+a_2 < 0$;

③ $a_1+a_2+a_3 < 1$;

④ a_4, a_5, a_6 non tous positifs.

Nous allons montrer que dans chacun de ces cas, il est possible de construire un indice d'agrégation tel que

$$\delta(h, h_2) \geq \delta(h, h_1) \geq \delta(h_1, h_2)$$

et $\delta(h_1, h_2) > \delta(h, h_1 \cup h_2)$ autrement dit :

$$(1-a_3)\,\delta(h_1,h_2) > (a_1-a_7)\,\delta(h,h_1) + (a_2+a_7)\,\delta(h,h_2) + a_4 f(h) + a_5 f(h_1) + a_6 f(h_2) \quad (3)$$

Considérons les quatre cas :

① $a_2+a_7 < 0$ ou $a_1 + a_7 < 0$.

On choisit $\delta(h,h_1) \geq \delta(h_1,h_2)$ et

$$\delta(h,h_2) > \text{Max}\left(\frac{(1-a_3)\,\delta(h_1, h_2) + (a_7 - a_1)\,\delta(h, h_1) - F}{a_2 + a_7},\ \delta(h, h_1)\right)$$

où $F = a_4 f(h) + a_5 f(h_1) + a_6 f(h_2)$.

② $a_1 + a_2 < 0$

On peut choisir

$$\delta(h,h_1) = \delta(h,h_2) > \text{Max}\left(\frac{(1-a_3)\,\delta(h_1, h_2) - F}{a_1 + a_2},\ \delta(h_1,h_2)\right)$$

③ $a_1 + a_2 + a_3 < 1$

On peut choisir

$$\delta(h,h_2) = \delta(h,h_1) + \varepsilon_1 = \delta(h_1,h_2) + \varepsilon_2 \text{ avec } \varepsilon_1 \text{ et } \varepsilon_2 \geq 0$$

On doit avoir

$$(a_1 - a_7)\,\delta(h,h_1) + (a_2 + a_7)\,(\delta(h,h_1) + \varepsilon_1) + F < (1-a_3)\,(\delta(h, h_1) + \varepsilon_1 - \varepsilon_2)$$

ou

$$(a_1 + a_2)\,\delta(h,h_1) + \varepsilon_1(a_2 + a_7) + F < (1 - a_3)\,\delta(h, h_1) + (1 - a_3)\,(\varepsilon_1 - \varepsilon_2)$$

$$C = \varepsilon_1(a_2 + a_7 + a_3 - 1) + \varepsilon_2(1 - a_3) + F < (1 - (a_1 + a_2 + a_3))\,\delta(h, h_1)$$

$D = 1-(a_1+a_2+a_3)$ étant strictement positif, il suffit de prendre $\delta(h,h_1) > \frac{C}{D}$ pour que l'inéquation (3) soit satisfaite.

④ a_4 , a_5 , a_6 non tous positifs.

On choisit $\delta(h,h_1) \geq \delta(h,h_2) \geq \delta(h_1,h_2)$ et l'inégalité (3) est satisfaite en prenant $f(h)$, $f(h_1)$ ou $f(h_2)$ suffisamment grand suivant que c'est a_1, a_2 ou a_3 qui sont strictement négatifs. Un tel choix est toujours possible puisqu'il y a $n^{(n-1)/2}$ couple de distances entre individus et $(n-1)(n-2)/2$ distances entre classes qui sont utilisés soit $(n-1)^2$ distances pour seulement $(n-1)(n-2)/2$ équations (voir §10). □

Il résulte de cette proposition qu'une condition nécessaire pour qu'il y ait inversion est que l'une au moins des conditions (a), (b), (c), (d), ne soit pas satisfaite.

6) Une condition nécessaire et suffisante pour l'existence d'inversions

Dans la proposition 2 nous n'avons pas considéré la condition B_1 suivante : ∀ δ satisfaisant à la formule de récurrence, la hiérarchie H qui s'en déduit par l'algorithme de la CAH, indicée par f définie par $a_1,\ldots,a_7$ a des inversions ; remarquons que B_1 n'est pas le contraire de B. Afin d'étudier cette condition, nous sommes amenés à distinguer le cas de la formule de récurrence générale où a_4, a_5 et a_6 ne sont pas tous nuls, du cas de la formule de Lance et Williams où $a_4 = a_5 = a_6 = 0$. Dans le cas de la formule générale on a le résultat suivant :

PROPOSITION 3

Une condition nécessaire et suffisante pour que tout δ satisfaisant à une formule de récurrence définie par des coefficients $a_1,\ldots,a_7$, induise par l'algorithme de la CAH, une hiérarchie indicée par f avec inversions est que les quatre conditions suivantes soient satisfaites :

(a^1) $\quad a_7 \leq -\text{Min}\ (a_1,a_2)$

(b^1) $\quad a_1 + a_2 \leq 0$

(c^1) $\quad a_1 + a_2 + a_3 \leq 1$

(d^1) $\quad a_4$, a_5 , a_6 tous négatifs non tous nuls.

Démonstration.

On dit que la condition A_1 est vraie si les relations (a^1), (b^1), (c^1) et (d^1) sont simultanément satisfaites.

Montrons d'abord la condition suffisante ; autrement dit que $A_1 \Longrightarrow B_1$.

Comme pour la proposition précédente on suppose que $\delta(h,h_1) \leq \delta(h,h_2)$; d'où :

$$\delta(h, h_1 \cup h_2) = (a_1 - a_7)\, \delta(h, h_1) + (a_2 + a_7)\, \delta(h, h_2) + a_3 \delta(h_1, h_2) + F.$$

où $F = a_4 f(h) + a_5 f(h_1) + a_6 f(h_2)$.

$(a^1) \Longrightarrow (a_2 + a_7) \leq 0 \Longrightarrow (a_2 + a_7)\, \delta(h, h_2) \leq (a_2 + a_7)\, \delta(h, h_1)$

d'où $\delta(h, h_1 \cup h_2) \leq (a_1 + a_2)\, \delta(h, h_1) + a_3 \delta(h_1, h_2) + F$

$(b^1) \Longrightarrow (a_1 + a_2)\, \delta(h, h_1) \leq (a_1 + a_2)\, \delta(h_1, h_2)$ car par construction $\delta(h, h_1) \geq \delta(h_1, h_2)$ d'où

$$\delta(h, h_1 \cup h_2) \leq (a_1 + a_2 + a_3)\, \delta(h_1, h_2) + F$$

$(c^1) \Longrightarrow \delta(h, h_1 \cup h_2) \leq \delta(h_1, h_2) + F$

$(d^1) \Longrightarrow \delta(h, h_1 \cup h_2) < \delta(h_1, h_2)$ qui prouve d'après la proposition 1 qu'il y a inversion.

Reste à montrer la condition nécessaire ; autrement dit que $B_1 \Longrightarrow A_1$; ou encore, que A_1 faux $\Longrightarrow B_1$ faux. Les quinze possibilités qui rendent A_1 faux sont couvertes par les quatre cas suivants :

① $a_7 > -\mathrm{Min}(a_1, a_2)$; ② $a_1 + a_2 > 0$; ③ $a_1 + a_2 + a_3 > 1$ et
④ $\exists$ un a_i $(i = 4, 5, 6)$ strictement positif (la possibilité $a_4 = a_5 = a_6 = 0$ est à exclure dans le cas de la formule générale)

Il faut montrer dans chacun de ces cas, que l'on peut construire un indice d'agrégation δ pour lequel la hiérarchie (H, f) n'admet pas d'inversion. Autrement dit, il faut montrer que pour tout h appartenant à la partition qui précède la formation du palier $h_1 \cup h_2$, il existe δ qui satisfasse dans chacun des quatre cas, aux inégalités suivantes :

$$\delta(h, h_2) \geq \delta(h, h_1) \geq \delta(h_1, h_2)$$

et $\delta(h, h_1 \cup h_2) \geq \delta(h_1, h_2)$, autrement dit :

$$(a_1 - a_7)\,\delta(h, h_1) + (a_2 + a_7)\,\delta(h, h_2) + F \geq (1 - a_3)\,\delta(h_1, h_2)$$

① $a_7 > -\text{Min}(a_1, a_2) \Longrightarrow a_2 + a_7 > 0$

On choisit $\delta(h,h_1) \geq \delta(h_1, h_2)$ et

$$\delta(h, h_2) \geq \text{Max}\left(\frac{(1-a_3)\,\delta(h_1, h_2) - (a_1 - a_7)\,\delta(h, h_1) - F}{a_2 + a_7}, \delta(h, h_1)\right)$$

Comme pour le cas ①, les cas de ②, ③ se traitent de façon analogue à la démonstration de la proposition 1 ; les inégalités sont en sens contraire mais les dénominateurs sont ici strictement positifs au lieu d'être strictement négatifs ; pour le cas ④, il suffit de prendre $f(h)$, $f(h_1)$ ou $f(h_2)$ suffisament grand suivant que c'est a_4, a_5 ou a_6 qui est strictement positif.

□

7) Une condition nécessaire à l'inexistence d'inversions

De la proposition 3 on déduit facilement la proposition suivante :

PROPOSITION 4

Une condition nécessaire pour qu'un indice d'agrégation δ satisfaisant à une formule de récurrence définie par des coefficients $a_1, \ldots, a_7$, induise par l'algorithme de la CAH une hiérarchie H indicée par f sans inversion est que l'une au moins des conditions suivantes soit satisfaite :

(a) $a_7 \geq -\text{Min}(a_1, a_2)$
(b) $a_1 + a_2 \geq 0$
(c) $a_1 + a_2 + a_3 \geq 1$
(d^2) a_4, a_5, a_6 tous positifs, non tous nuls.

On a un résultat analogue dans le cas de la formule de Lance et Williams, en se restreignant aux conditions (a), (b) et (c).

Comme nous l'avons vu dans la proposition 2 chacune de ces conditions n'est pas suffisante pour assurer l'inexistence d'inversions, il faut qu'elles soient toutes satisfaites simultanément.

8) Cas de la formule de Lance et Williams ($a_4 = a_5 = a_6 = 0$)

PROPOSITION 5

Une condition suffisante pour que la condition B_1 (donnée en 6)) soit satisfaite est que les 3 conditions suivantes soient satisfaites :

(a^2) $a_7 < -\text{Min}(a_1, a_2)$
(b^2) $a_1 + a_2 < 0$
(c^2) $a_1 + a_2 + a_3 < 1$

où deux au plus de ces inégalités peuvent être larges.

La démonstration de cette proposition est tout à fait analogue à celle de la proposition 3.

9) Problèmes d'inversions dans les algorithmes accélérés pour les grands tableaux

Parmi toutes les distances qu'il est nécessaire de calculer pour la construction des classes d'une hiérarchie ascendante, au cours de l'algorithme général, seules sont retenues les plus petites. Partant de cette idée, il est naturel d'imaginer un algorithme dans lequel les seules distances utiles soient calculées. Plus modestement, il s'agit de calculer le moins de distances inutiles possible. La question se posait alors de savoir si la hiérarchie ainsi obtenue était toujours la même que la hiérachie que l'on aurait obtenu par l'algorithme général usuel. Bruynooghe (1978) a alors montré qu'il n'y a pas de déformation de la hiérarchie si une propriété dite de "réductibilité" est satisfaite.

9.1) Définition de la propriété de réductibilité

On suppose donné un nombre ρ strictement positif. Soit P_h la partition qui dans le cours de l'algorithme précède la formation de $h = h_1 \cup h_2$. On pose :

$B_\rho(h) = \{h_i \in P_h \;/\; \delta(h_i, h_1 \cup h_2) < \rho \text{ avec } h_i \neq h_1 \text{ et } h_i \neq h_2\}$, (voir figure 3).

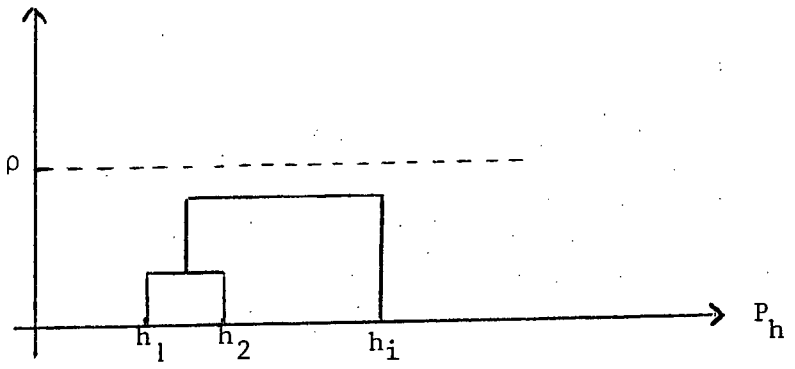


Figure 3

On dit que δ satisfait à la propriété de réductibilité si $B_\rho(h_1 \cup h_2) \subset$ $\subset B_\rho(h_1) \cup B_\rho(h_2)$ où $h = h_1 \cup h_2$, h_1 et h_2 sont des paliers de la hiérarchie H avec $\delta(h_1, h_2) \leq \rho$.

Cette condition peut également s'exprimer sous les deux formes suivantes :

δ satisfait à la propriété de réductibilité si pour tout $h_i \in P_h$ avec $h = h_1 \cup h_2$, $h_i \neq h_1$ et $h_i \neq h_2$, on a pour tout $\rho > 0$:

$$\left.\begin{array}{l}\delta(h_i, h_1) \geq \rho \\ \delta(h_i, h_2) \geq \rho \\ \delta(h_1, h_2) \leq \rho\end{array}\right\} \Longrightarrow \delta(h_i, h_1 \cup h_2) \geq \rho \qquad (4)$$

Autrement dit les paliers de P_h qui sont à une "distance" de h_1 et de h_2 supérieure à ρ doivent être à une distance de $h_1 \cup h_2$ supérieure à ρ si $\delta(h_1, h_2) \leq \rho$.

On peut aussi dire que δ satisfait à la propriété de réductibilité si :

$$\left.\begin{array}{l}\delta(h_i, h_1 \cup h_2) < \rho \\ \delta(h_1, h_2) \leq \rho\end{array}\right\} \Longrightarrow \delta(h_i, h_1) < \rho \text{ ou } \delta(h_i, h_2) < \rho \,. \qquad (5)$$

Ce qui signifie que les paliers qui sont à une "distance" de $h = h_1 \cup h_2$ inférieure à ρ sont nécessairement à une "distance" inférieure à ρ de h_1 ou de h_2 quand $\delta(h_1, h_2) \leq \rho$.

9.2) Lien entre inversion et réductibilité

PROPOSITION 6

Une condition nécessaire et suffisante pour qu'une hiérarchie construite par l'algorithme de la CAH à l'aide de l'indice d'agrégation δ soit sans inversions est que δ satisfasse à la propriété de réductibilité.

Démonstration

Montrons d'abord la condition suffisante . Nous allons montrer que si δ satisfait à la propriété de réductibilité alors la condition (3) de la proposition 1 est satisfaite. Cette condition dite "globale" s'énonce sous la forme suivante (voir figure 2) :

$$\left.\begin{array}{l}\text{pour tout } h_j \in P_h \text{ tels que } h_j \neq h_1,\ h_j \neq h_2,\ h = h_1 \cup h_2 \in H : \\ \text{on a } \delta(h_j, h) \geq \delta(h_1, h_2).\end{array}\right\} \quad (6)$$

Or dire que $h_j \in P_h$ avec $h_j \neq h_1$, $h_j \neq h_2$ et $h = h_1 \cup h_2$ implique que :

$\delta(h_j, h_1) \geq \delta(h_1, h_2)$ et $\delta(h_j, h_2) \geq \delta(h_1, h_2)$ car P_h est la partition qui dans le cours de l'algorithme précède la formation de $h = h_1 \cup h_2$. On déduit de ces deux inégalités l'existence de ρ :

$\delta(h_j, h_1) \geq \rho \geq \delta(h_1, h_2)$ et $\delta(h_j, h_2) \geq \rho \geq \delta(h_1, h_2)$ d'où $\delta(h_j, h_1) \geq \rho$, $\delta(h_j, h_2) \geq \rho$ et $\delta(h_1, h_2) \leq \rho$.

En appliquant la propriété de réductibilité sous la forme (4), on en déduit que $\delta(h_j, h_1 \cup h_2) \geq \rho \geq \delta(h_1, h_2)$.

La relation (6) est donc satisfaite ; d'après la proposition 1, on a bien une hiérarchie indicée.

Montrons maintenant la condition nécessaire ; il faut montrer que s'il n'y a pas inversion il y a réductibilité ; pour cela, nous allons montrer que s'il n'y a pas réductibilité il y a inversion. Si la réductibilité n'est pas satisfaite, il existe $\rho > 0$ tel que pour tout $h_i \in P_{h_1 \cup h_2}$ différent de h_1 et h_2 on ait :

$$\left.\begin{array}{l}\delta(h_i, h_1) \geq \rho \\ \delta(h_i, h_2) \geq \rho \\ \delta(h_1, h_2) \leq \rho\end{array}\right\} \Rightarrow \delta(h_i, h_1 \cup h_2) < \rho$$

Choisissons $\rho = \delta(h_1, h_2)$; à l'étape qui suit la formation de $h = h_1 \cup h_2$ il existe $h_i \in P_h$ tel que $\delta(h_i, h) < \rho = \delta(h_1, h_2)$; il y a donc une inversion d'après la proposition 1.

□

PROPOSITION 7

Une condition suffisante pour qu'un indice d'agrégation δ satisfaisant à la formule de récurrence soit réductible est que les conditions (a), (b), (c) et (d) soient satisfaites. Une condition nécessaire étant que l'une au moins des conditions (a), (b), (c) et (d^2) le soit.

Démonstration

Montrons d'abord la condition suffisante. De la proposition 2 on déduit le fait que si les conditions (a), (b), (c) et (d) sont satisfaites alors il n'y a pas d'inversions ; d'après la proposition 6 la réductibilité est alors satisfaite.

La condition nécessaire se déduit de la proposition 4 qui prouve que pour ne pas avoir d'inversions et donc pour avoir la réductibilité (d'après la proposition 6) il faut que l'une au moins des conditions (a), (b), (c) et (d^2) soient satisfaites.

□

10) Etude et extension des formules de récurrence classiques : construction adaptive d'indices d'agrégation.

Etant donnée une formule de récurrence on peut dire pour simplifier que soit les conditions (a), (b), (c), (d) de la proposition 2 sont satisfaites, alors il n'y a pas d'inversions (proposition 2), soit aucune ne l'est, alors il y a inversion (proposition 3), soit elles sont partiellement satisfaites alors on ne peut se prononcer.

En annexe 2 nous donnons une série de formules de récurrence classiques suivies d'un tableau donnant les propriétés d'inversions et de réductibilité des indices d'agrégation correspondants. On peut étendre la famille des indices d'agrégation classiques en calculant les a_i à l'aide de la formule de récurrence à partir d'un exemple. Pratiquement l'utilisateur propose, sur un exemple issu de ses données, la hiérarchie H désirée en précisant la hauteur de chacun des paliers. A partir de cette hiérarchie on peut calculer $\delta(h_i, h_j)$ où h_i et h_j sont deux paliers quelconques de H en posant :

$$\delta(h_i, h_j) = \{\text{la hauteur du plus bas palier contenant } h_i \text{ et } h_j\}.$$

A l'aide de ces quantités et de la formule de récurrence générale on obtient un système d'équations avec 7 inconnues (les a_i pour i = 1, ..., 7) comportant $\frac{(n-1)(n-2)}{2}$ égalités (n est la taille de la population) ; en effet :

P_{h_i} étant la partition qui précède la formation du $i^{ème}$ palier h_i, on utilise n-2 équation de récurrence pour construire h_2 ; s'il faut n-m nouvelles équations pour construire h_m il en faut n-(m+1) pour construire h_{m+1} (si h_{m+1} contient h_m il en faut n-m-1 sinon n-m-2+1). Il faut donc utiliser en tout $(n-2) + (n-3) + \ldots + 1 = \frac{(n-1)(n-2)}{2}$ équations pour construire la hiérarchie.

Il faut trouver les a_i qui satisfont au mieux ces équations et les contraintes données par la proposition 2 ou la proposition 4 (par exemple, les conditions (a), (b), (c), (d) de la proposition 2 pour assurer l'inexistence d'inversions). En règle générale il y a plus d'équations que d'inconnues ; la solution qui est de type moindre carrés sous contraintes ne satisfait pas nécéssairement toutes les égalités : des inversions peuvent donc apparaître ; pour palier à cet inconvénient, l'utilisateur pourra remettre en question ces choix de façon inter-active (en modifiant par exemple les facteurs des paliers correspondant à ces inversions) jusqu'à obtenir les coefficients correspondant au mieux à son désir.

Exemple

. Les données : quatre points du plan dont les distances $d(w_i, w_j)$ sont précisées sur la figure 4

	w_1	w_2	w_3	w_4
w_1	0	1	3	2
w_2	1	0	2	3
w_3	3	2	0	2
w_4	2	3	2	0

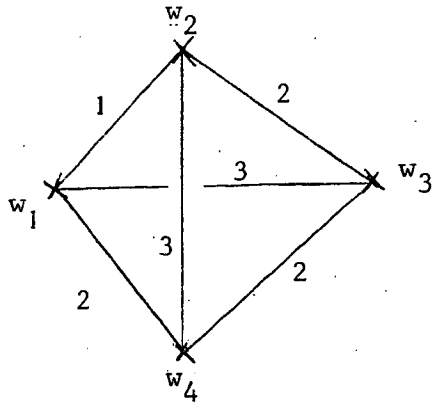
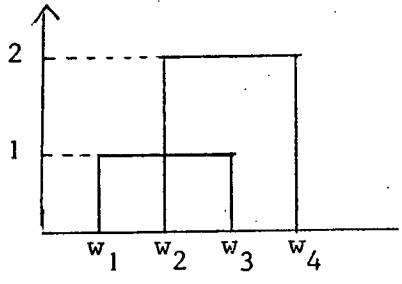


Figure 4

. La hiérarchie désirée est donnée figure 5



$h_1 = \{w_1\} \cup \{w_2\}$

$h_2 = h_1 \cup w_3$

$h_4 = h_2 \cup w_4$

Figure 5

. Les équations à résoudre

Etant donné le faible nombre d'individus on se place dans le cas de la formule de Lance et Williams ($a_4 = a_5 = a_6 = 0$) ; on a :

$$\delta(w_i, w_j) = d(w_i, w_j)$$

Avec i = 3, 4 on obtient 2 équations du type suivant :

. $\delta(w_i, h_1) = a_1\delta(w_i, w_1) + a_2\delta(w_i, w_2) + a_3\delta(w_1, w_2) + a_7|\delta(w_i, w_2) - \delta(w_i, w_1)$

enfin :. $\delta(w_4, h_2) = a_1\delta(w_4, h_1) + a_2\delta(w_4, w_3) + a_3\delta(h_1, w_3) + a_7|\delta(w_4, w_3) - \delta(w_4, h_1)$

Avec les valeurs prises par δ on obtient les équations suivantes :

$$1 = 3a_1 + 2a_2 + a_3 + a_7$$
$$2 = 2a_1 + 3a_2 + a_3 + a_7$$
$$2 = 2a_1 + 2a_2 + a_3$$

Avec les contraintes :

$$a_7 \geq -\text{Min}(a_1, a_2)$$
$$a_1 + a_2 \geq 0$$
$$a_1 + a_2 + a_3 \geq 1$$

Si les deux dernières inégalités sont satisfaites, la condition nécessaire l'est également (proposition 4). Une solution assurant une symétrie entre a_1 et a_2 est donnée par

$$a_1 = -a_2 = a_7 = -\frac{1}{2} \qquad a_3 = 2$$

Comme a_1 et a_2 ne sont pas égaux, on peut obtenir plusieurs hiérarchies satisfaisant à cette formule de récurrence. On obtient une hiérarchie unique en associant à a_1 la classe de plus petit indice. En la construisant de cette façon on obtient la hiérarchie désirée par l'utilisateur sur les quatre points de l'exemple.

Conclusion

Parmi l'ensemble des résultats donnés on peut retenir les règles suivantes : si les conditions (a), (b), (c), (d) sont satisfaites on est assuré de l'inexistence d'inversions, si aucune ne l'est il y a inversion, il est nécessaire que l'une au moins soit vraie pour ne pas avoir d'inversions. Dans tous les domaines où les utilisateurs ont une idée précise des classes désirées à partir d'exemples particuliers (traitement d'image, par exemple), les propriétés données débouchent sur des programmes de génération automatique de hiérarchies à partir d'exemples.

Afin d'enrichir les possibilités d'adéquation de la formule de récurrence à l'exemple proposé par l'utilisateur, il serait intéressant d'étudier la formule de récurrence générale suivante :

$$\begin{aligned}\delta(h, h_1 \cup h_2) = {} & a_1\, \delta(h, h_1) + a_2\, \delta(h, h_2) + a_3\, \delta(h_1, h_2) \\ & + a_4\, f(h) + a_5\, f(h_1) + a_6\, f(h_2) \\ & + a_7\, |f(h) - f(h_1)| + a_8\, |f(h) - f(h_2)| + a_9\, |f(h_1) - f(h_2)| \\ & + a_{10}\, |\delta(h, h_1) - \delta(h, h_2)|\end{aligned}$$

Avec $a_i = F_i\,(p(h), p(h_1), p(h_2))$, où $p(h_i)$ est un poids associé au palier h_i.

ANNEXE 1

L'algorithme général de la classification ascendante hiérarchique

Cet algorithme (dit de la C.A.H.) consiste à construire à l'aide de l'indice d'agrégation δ choisi, une suite de partitions de moins en moins fines dont les classes forment la hiérarchie H cherchée. Il s'énonce de la façon suivante :

(1) Partir de la partition P_o dont les classes sont réduites à un élément.

(2) Construire une nouvelle partition en réunissant les deux classes de la partition précédente qui minimisent δ.

(3) Recommencer le procédé en 2 jusqu'à ce que toutes les classes soient réunies en une seule.

. Si à l'étape (2) il y a plus d'un couple de classes qui minimise δ, on en choisit un au hasard ; il n'y a donc pas toujours unicité de la hiérarchie obtenue. On remarque aussi que la hiérarchie ainsi obtenue est nécessairement binaire.

ANNEXE 2

Valeur des coefficients de la formule de récurence pour quelques indices d'agrégation.

. Indice du saut minimum :

$$\delta_1(h_1,h_2) = \text{Min } \{d(w_1,w_2)/w_1 \in h_1, w_2 \in h_2\}$$

$$a_1 = a_2 = \frac{1}{2} \qquad a_3 = a_4 = a_5 = a_6 = 0 \qquad a_7 = -\frac{1}{2}$$

. Indice du saut maximum :

$$\delta_2(h_1,h_2) = \text{Max } \{d(w_1,w_2)/w_1 \in h_1, w_2 \in h_2\}$$

$$a_1 = a_2 = \frac{1}{2} \qquad a_3 = a_4 = a_5 = a_6 = 0 \qquad a_7 = \frac{1}{2}$$

. Indice d'agrégation de la moyenne des distances

$$\delta_3(h_1,h_2) = \frac{1}{|h_1||h_2|} \sum_{\substack{w_i \in h_1 \\ w_j \in h_2}} d(w_i,w_j) \quad \text{où } |h_i| = \text{card } h_i$$

$$a_1 = \frac{|h_1|}{|h_1| + |h|}, \quad a_2 = \frac{|h_2|}{|h_2| + |h|} \qquad a_i = 0 \text{ pour } i > 2$$

. Indice de la distance des centres de gravité

$$\delta_4(h_1, h_2) = d(G(h_1), G(h_2))$$

$$a_1 = \frac{p(h_1)}{p(h_1) + p(h)} \qquad a_2 = \frac{p(h_2)}{p(h_2) + p(h)} \qquad a_3 = \frac{-p(h_1)\ p(h_2)}{(p(h_1) + p(h_2))^2}$$

$$a_i = 0 \text{ pour } i > 3$$

où $p(h) = \sum_{w \in h} p(h)$ est le poids associé à h.

. Indice de l'inertie

$\delta_5 (h_1,h_2) = I (h_1 \cup h_2)$ où

$I(h) = \sum_{w \in h} p(w)\, d(w,G(h))$; $G(h)$ est le barycentre de h ; $p(w)$ est un poids associé à chaque individu : $p(h_1 \cup h_2) = p(h_1)+p(h_2)$.

$$a_1 = \frac{p(h)+p(h_1)}{T} \qquad a_2 = \frac{p(h)+p(h_2)}{T} \qquad a_3 = \frac{p(h_1)+p(h_2)}{T}$$

$$a_4 = -\frac{p(h_1)}{T} \qquad a_5 = -\frac{p(h_2)}{T} \qquad a_6 = -\frac{p(h)}{T} \qquad a_7 = 0$$

où $T = p(h) + p(h_1) + p(h_2)$

. Indice de la variance

$$\delta_6(h_1,h_2) = \text{var}\,(h_1 \cup h_2) = \frac{1}{p(h_1) + p(h_2)}\, I(h_1 \cup h_2).$$

$$a_i = \left[\frac{p(h) + p(h_i)}{T}\right]^2 \quad \text{pour } i = 1, 2.$$

$$a_3 = \left[\frac{p(h_1) + p(h_2)}{T}\right]^2 \qquad a_4 = -\frac{p(h_1)^2}{T^2} \qquad a_5 = \frac{-\,p(h_2)^2}{T^2} \qquad a_6 = \frac{-\,p(h)^2}{T^2}$$

. Indice de l'augmentation d'inertie (Ward 1963)

$$\delta_7 (h_1,h_2) = I(h_1 \cup h_2) - I(h_1) - I(h_2) = \frac{p(h_1)\, p(h_2)}{p(h_1)+p(h_2)}\, d(G(h_1), G(h_2))^*$$

$$a_1 = \frac{p(h)+p(h_1)}{T} \qquad a_2 = \frac{p(h)+p(h_2)}{T} \qquad a_3 = \frac{-p(h)}{T}$$

$$a_4 = a_5 = a_6 = a_7 = 0$$

* Si d est une distance euclidienne.

. Indice de l'augmentation pondérée de variance

$$\delta_8(h_1, h_2) = \text{var}\,(h_1 \cup h_2) - \frac{p(h_1)}{p(h_1 \cup h_2)}\ \text{var}\,h_1 - \frac{p(h_2)}{p(h_1 \cup h_2)}\ \text{var}\,h_2.$$

$$a_i = \left[\frac{p(h) + p(h_i)}{T}\right]^2 \quad \text{pour } i = 1, 2 \qquad a_3 = -\frac{p(h)((p(h_1) + p(h_2))}{T^2}$$

$a_i = 0$ pour $i > 3$.

Propriétés d'inversion et de réductibilité

Dans le tableau suivant les conditions (a), (b), (c), (d) de la proposition 2 sont notées a, b, c, d ; (non a) est noté $\bar{a}$.

Indices d'agrégation	Conditions satisfaites	Propriétés
δ_1	a b c d	pas d'inversions*, réductibilité*
δ_2	a b c d	pas d'inversions*, réductibilité*
δ_3	a b c d	pas d'inversions*, réductibilité**
δ_4	a b $\bar{c}$ d	possibilité d'être sans inversions***
δ_5	a b c $\bar{d}$	pas d'inversions°, réductibilité**
δ_6	a b c $\bar{d}$	possibilité d'être sans inversions***
δ_7	a b c d	pas d'inversions*, réductibilité**
δ_8	a b $\bar{c}$ d	pas d'inversions°, réductibilité**

* D'après la proposition 2.

** D'après la proposition 6.

*** D'après la proposition 4.

° Car $I(h_1 \cup h_2) \geq I(h_1) + I(h_2)$ (Huygens) d'où $\delta(h_1, h_2) \geq \text{Max}\,(f(h_1), f(h_2))$ en supposant bien sûr que d est une distance euclidienne.

BIBLIOGRAPHIE

[1] BATAGELJ V. "Note on ultrametric hierarchical clustering algorithms" Psychometrica, Vol. 46 n° 3 (1981).

[2] DIDAY E., LEMAIRE J., POUJET J., TESTU F., "Eléments d'analyse des données", Dunod, (1982)

[3] P. DUMICETIERE, "Les méthodes de classification numérique" Revue de Statistisques Appliquées, Volume 18, n° 4 p. 5-25 (1970)

[4] BRUYNOOGHE M., "Classification ascendante hiérarchique de grands ensembles de données : un algorithme rapide fondé sur la construction des voisinages réductibles", Cahiers d'Analyse des Données, Vol. III, n° 1, (1978).

[5] JAMBU M., "Classification automatique pour l'analyse des données", Dunod. (1978).

[6] LANCE G.C., WILLIAMS W.T., "A general theory of classification sorting", Computer Journal 9.10 and Computer Journal 10.3 (1967).

[7] MILLIGAN G., "Ultrametric hierarchical clustering algorithms", Psychometrica 44,3. (1979).

Imprimé en France

par

l'Institut National de Recherche en Informatique et en Automatique